**Great Learning**

**Capstone Project – Final Report (Milestone 2)**

**Pneumonia Detection Challenge**

**Group:** **CV Group 4**

**Authors:**

1. **Ajay Devnani**
2. **Manojkumar Sivaraman**
3. **Naveen Kumar Kaginelli**
4. **Sumeet Kumar**
5. **Umesh Singh Kushwaha**

**Abstract**

Pneumonia is an infection in one or both lungs. Bacteria, viruses, and fungi cause it and the infection causes inflammation in the air sacs in your lungs, which are called alveoli. The alveoli fill with fluid or pus, making it difficult to breathe. Typically, X-ray helps doctors to look for signs of inflammation or opacities in chest which when present can indicate the Pneumonia infections. Since Pneumonia accounts for over 15% of all deaths of children under 5 years old internationally, it is crucial to identify and react swiftly if there are any infections identified.

**Business Perspective**: As the detection and reaction time is vital and the infection is detected using the X-ray Images, Image processing techniques can be leveraged from the emerging AI technology on these images to predict the presence of opacities. Powerful AI techniques can unlock clinically relevant information hidden in the massive amount of data, which in turn can assist clinical decision making. This will also assist physicians to make better clinical decisions or even replace human judgement in certain functional areas of healthcare (e.g., radiology).

To achieve the goal, techniques such as instance segmentation (Mask RCNN) and semantic segmentation are used to create the Pneumonia prediction model which can predict Pneumonia and also the locations of inflammation (lung opacities) in lungs.

1. **Problem Statement**

The problem is about detecting lung inflammations (opacities) corresponding diagnosis of Pneumonia on chest radiographs (CXR). Pneumonia usually manifests as an area or areas (inflammation areas) of increased opacity on CXR.

Tissues with sparse material, such as lungs which are full of air, do not absorb the X-rays and appear black in the image. Dense tissues such as bones absorb X-rays and appear white in the image. In the data, some of these such area labeled as “Not Normal No Lung Opacity”.

The “Not Normal No Lung Opacity” category indicates that, while pneumonia determined not to be present, there could be nonetheless some type of abnormality on the radiograph. And oftentimes this finding may mimic the appearance of true pneumonia. A radiograph may contain one or more than one locations for any possible Pneumonia case.

**Findings and Implications:**

* Lung opacities in chest radiographs could be due to various abnormal conditions including pneumonia. Therefore, all opacities may not be necessarily due to Pneumonia.
* A radiograph may contain one or more than one lung inflammation areas.
* A number of factors such as positioning of the patient and depth of inspiration can alter the appearance of the chest radiograph.

The diagnosis of pneumonia on chest radiograph (CXR) is quite complicated because of a number of other conditions in the lungs such as fluid overload (pulmonary edema), bleeding, volume loss (atelectasis or collapse), lung cancer, or post-radiation or surgical changes.

When available, comparison of CXRs of the patient taken at different time points and correlation with clinical symptoms and history are helpful in making the diagnosis.

1. **Data and Algorithms Used**

The goal is to predict whether a given patient is suffering from Pneumonia or not. There are 2 related tasks,

* The radiograph images have to analyzed for potential lung inflammation areas (lung opacities).
* Predict whether the lung inflammation areas attribute to pneumonia or not.

That means our model/algorithm (the solution) should predict potential areas for lung inflammation and categorize the patient to be having Pneumonia or not having Pneumonia based on the prominence of the predicted inflammation areas.

The given data set includes 26,684 training samples and 3000 testing samples. The Training data set contains a chest radio graph (DICOM file) along with information regarding the label (i.e. diagnosed as pneumonia or not) and lung inflammation areas for each patient. Whereas, he Test set contains only the radiographs, no target data is given.

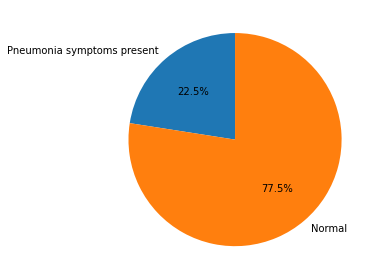
The data is spread across different files and folders. The details are as given below,

* **stage\_2\_train\_images:** It contains a set of raw medical images (DICOM files) for training models. The DICOM files contain a combination of header metadata as well as underlying raw image arrays for pixel data.
* **stage\_2\_test\_images:** It contains a set of raw medical images (DICOM files) for testing the model. The file contains a combination of header metadata as well as underlying raw image arrays for pixel data**.**
* **stage\_2\_train\_labels.csv:** This CSV file contains detailed information about the labels (Patient ID, bounding boxes for lung opacity and target 1 or 0 indicate the presence of abnormality i.e. Pneumonia)
* **stage\_2\_detailed\_class\_info.csv:** This CSV file contains information regarding three possible classes in the data - namely normal, lung opacity and no lung opacity (not normal).
* **DICOM files:** The original medical images are stored in a special format called DICOM files (\*.dcm). It contains a combination of header metadata as well as underlying raw image arrays for pixel data.

**Exploratory data analysis (EDA):**

1. **Distribution of Pneumonia Vs Non-Pneumonia:**

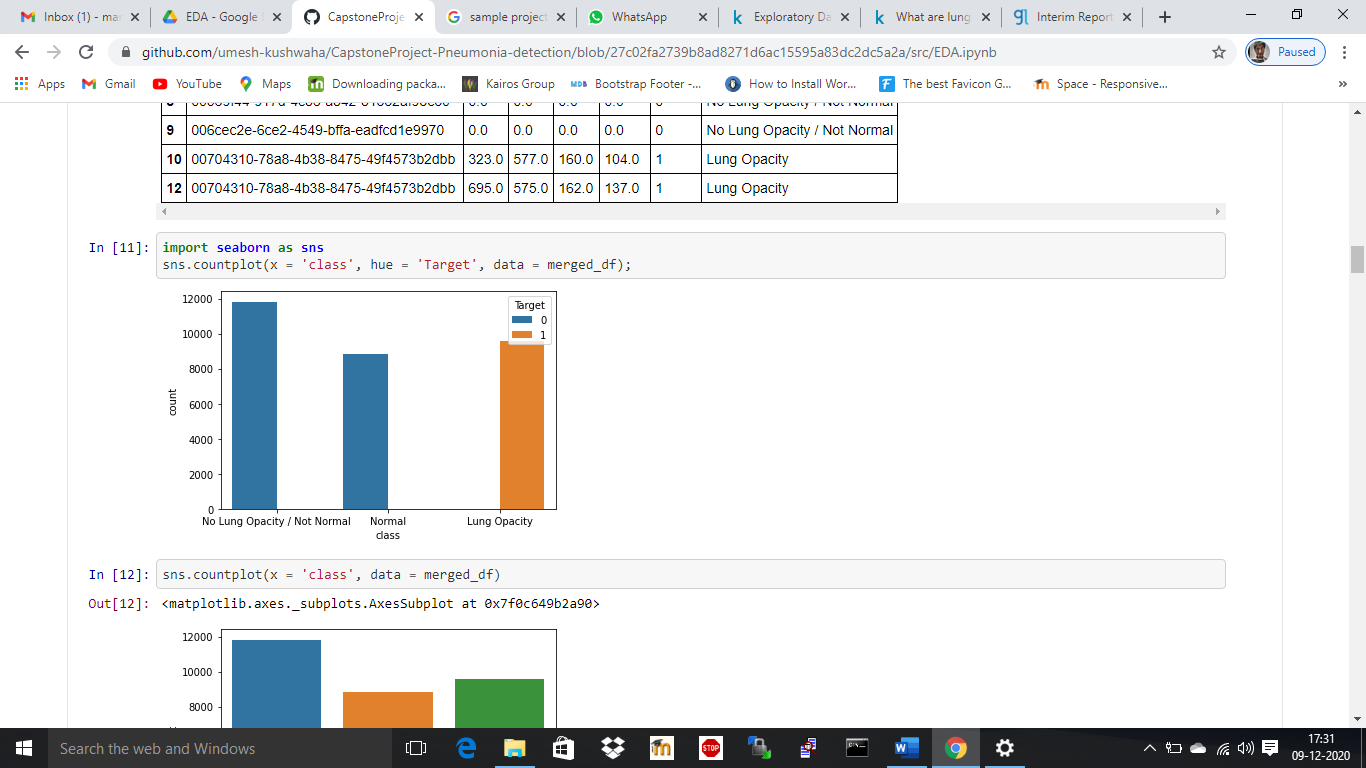
* 33.2% (8,851) patients are normal, do not have any lung related abnormalities
* 22.5% (6012) patients have *lung opacities* which attributes to pneumonia.
* 44.3% (11,821) patients do not have pneumonia but are not normal possibly due to other lung ailments.
* Hence, **22.5% of patients are suffering from pneumonia** and the remaining **77.5% are pneumonia negative**.

****

**Action Taken:** Data is not balanced, to balance the data augmentation technique is used during data generation.

1. **Distribution of 3 different classes data with target:**

* The count of patients with *No Lung opacities/ Not normal* is higher than the pneumonic or normal patients
* the count of normal class is less than other 2 classes indicating that the data has a greater number of Ill health patients

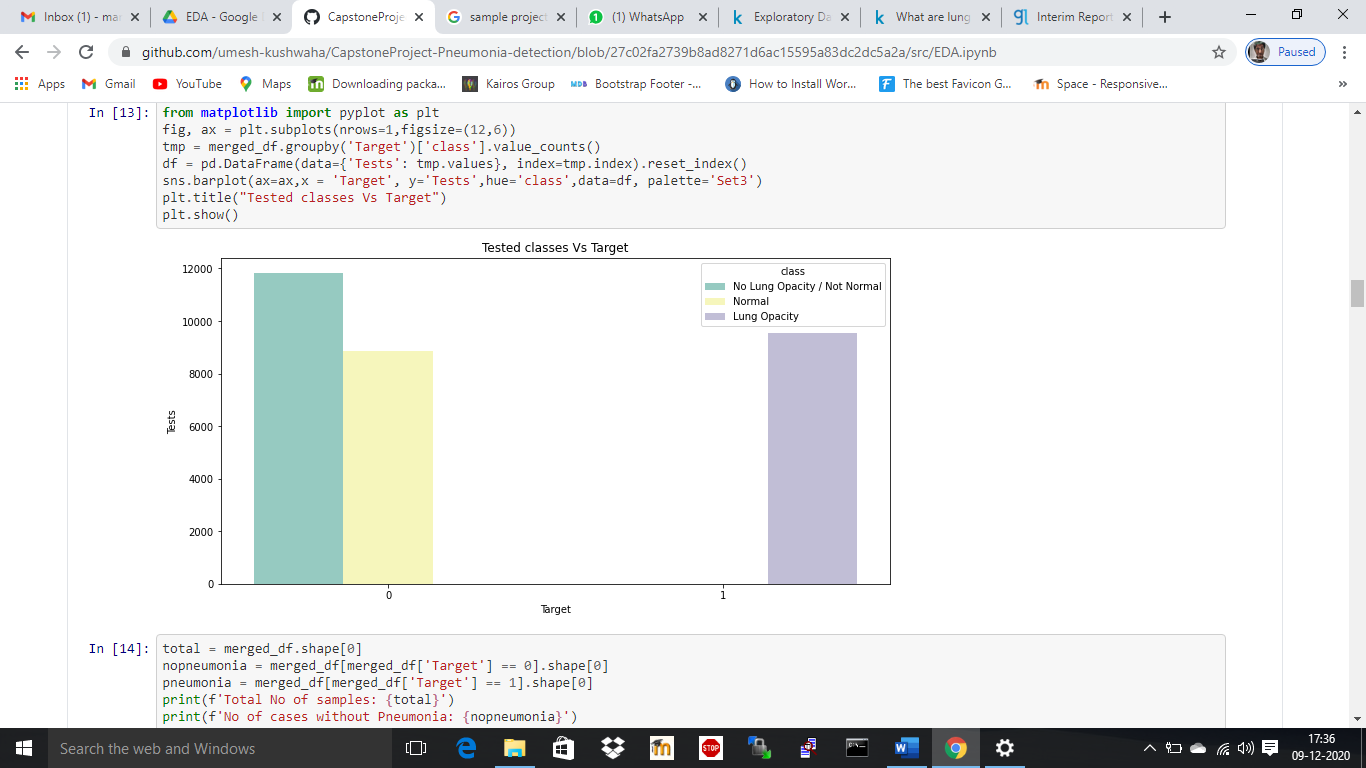


**Action Taken:** No Lung Opacity / Not normal class corresponds to pneumonia but other possible lung ailments. Therefore, this class is considered as no pneumonia i.e. normal patients.

1. **Distribution of target (Pneumonia & Non-Pneumonia) with 3 different classes:**

The same can be observed when plotting the count of Target values segregating the classes.

* Patients with *No Lung Opacity/ Not normal* observations are more than there other 2 classes



**Action Taken:** No Lung Opacity / Not normal class corresponds to pneumonia but other possible lung ailments. Therefore, this class is considered as no pneumonia i.e. normal patients.

1. **Train set:**

* 26,684 images are available in the training set are unique (equal to unique patient IDs).

1. **Bounding box:**

* Out of 26,684 images available in the training set, 2614 images have only 1 bounding box, 3,266 images have 2 bounding boxes, 119 images have 3 boxes, 13 images have 4 boxes and 20672 images do not contain lung opacities.
* 3,398 patients have more than 1 bounding box.
* If any patient has at least one lung opacity area, then the patient is considered to have pneumonia.

|  |  |
| --- | --- |
| **No of Occurrences** | **Count of the Patient ID** |
| 0 | 20672 |
| 1 | 2614 |
| 2 | 3266 |
| 3 | 119 |
| 4 | 13 |

1. **Data characteristics:**

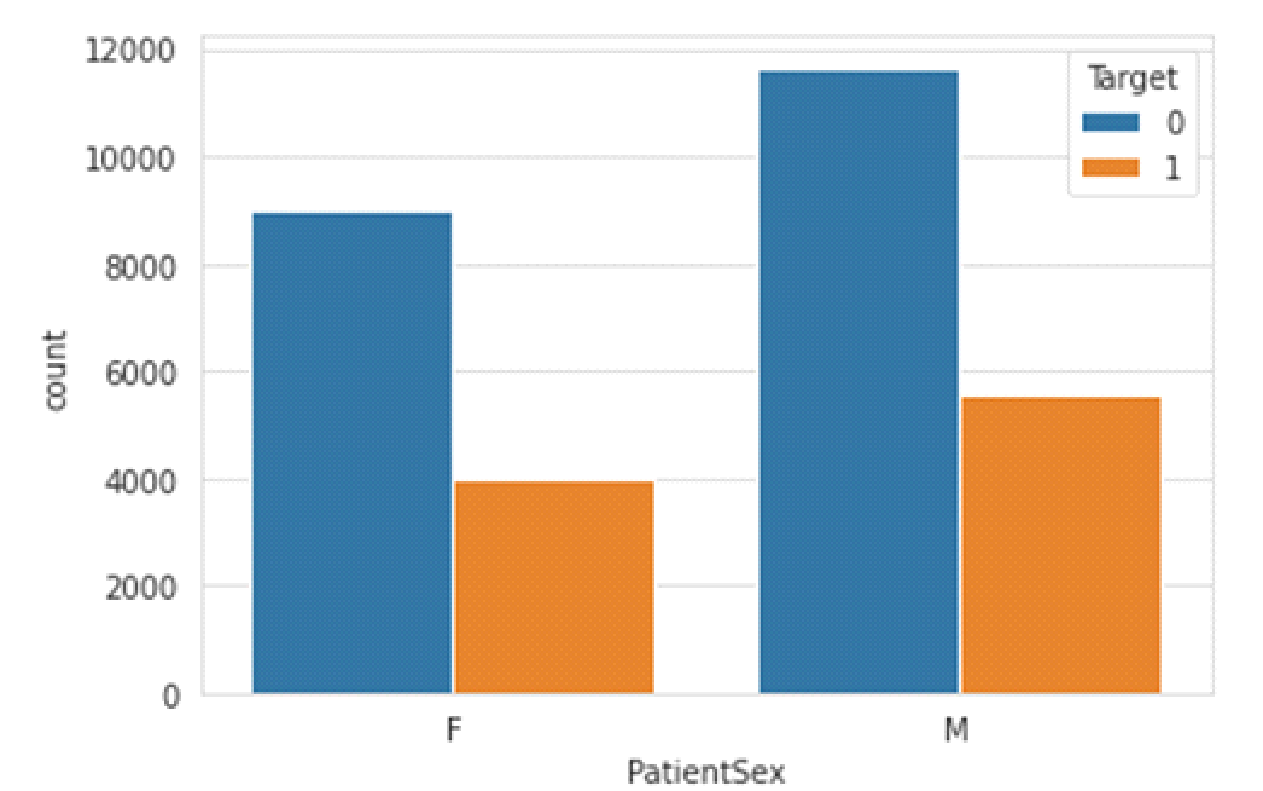
* We have different parameters or characteristics of available information – patient age, sex, body part examined, view position, rows and columns, pixel spacing, etc.

1. **Correlation:**

* We have observed that ‘Target’ and ‘View Position’ have a higher correlation and stand at 0.42.

1. **Gender mix: Distribution Gender Vs Target information**

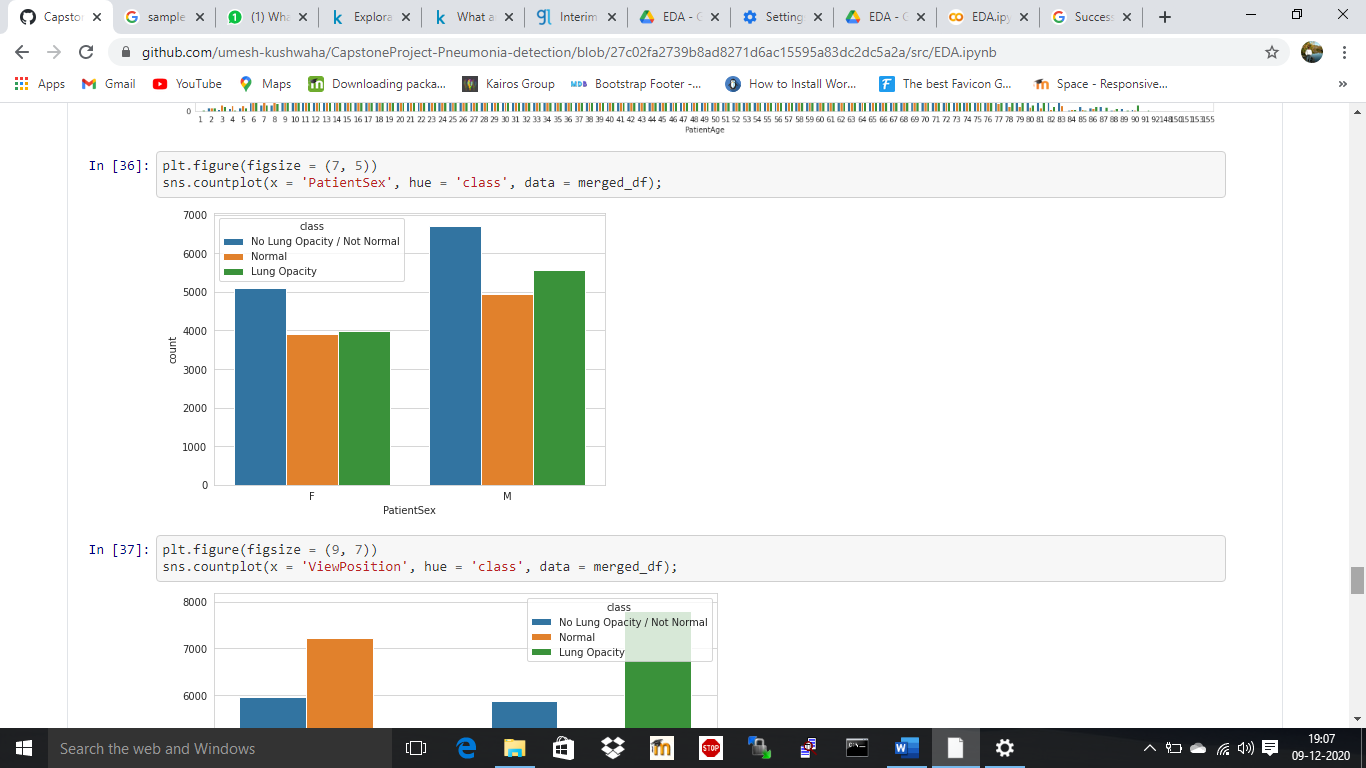
* Looking at the high pillars, there are more test samples (radiographs) for males than females.
* This could be due to, Men consume more alcohol, work outdoor more than women, smoke.
* Out of total 9,555 cases of Lung opacity, ~60% is male and rest 40% is female.
* Approximately, one third of the total cases are diagnosed as pneumonia for both the genders.



**Action Taken**: To balance the data augmentation with random shuffling is used.

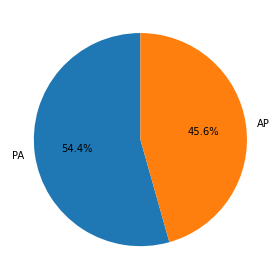
1. **Gender mix: Distribution Gender Vs class information**

* No opacity but Not Normal cases constitute higher number indicating that the patients could be suffering from other lung related illness but pneumonia.
* Males who are diagnosed for lung opacities (Pneumonia) are slightly high in number comparatively.

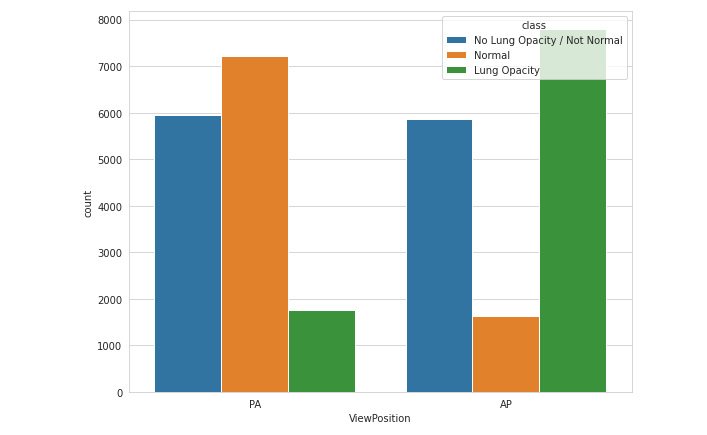


**Action Taken**: To balance the data augmentation with random shuffling is used.

1. **Distribution of View position:**



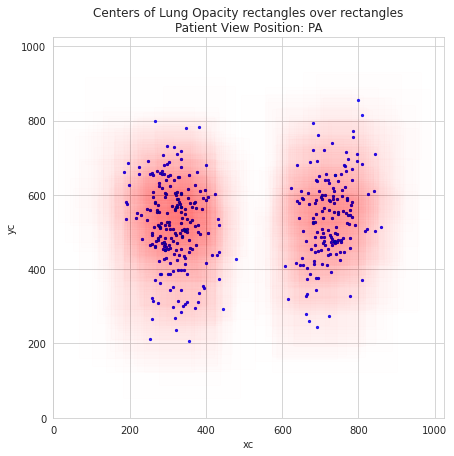
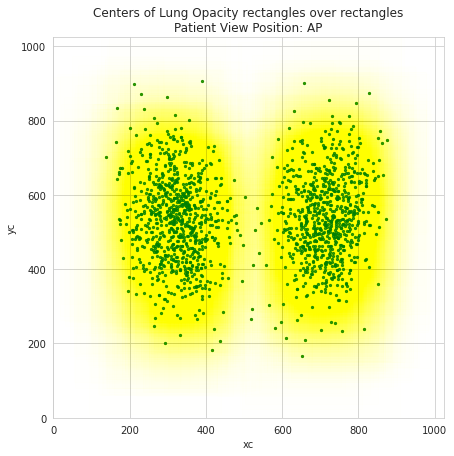
* Radiographs images with PA position are considered to be of good standard in medical profession. The PA amounts to 54.4% of total radiograph images. Therefore, data augmentation with random shuffling to balance the data to spread the impact of AP evenly.



* Evidently **PA position pointing considerably less lung opacities** than the AP position. Whereas the *no opacity but not normal* class seems to be the same in both the position.
* Scanning in AP position is usually done when the patient is not able to standup or for some reason the frontal scanning is not possible.
* This means a patient who has undergone scanning in AP position is more likely to be diagnosed for Pneumonia.
* If the same patient undergoes scanning in PA position, chances are that the results might come out normal.

**Action Taken**: Data augmentation with random sampling is used to balance the data. However, this is an important insight to be noted and the current solution not making use of this explicitly. It’s mentioned as part of implication and limitations section.

* In the below chart, concentration of Lung opacities for AP is larger, whereas for the PA cases, it’s less.



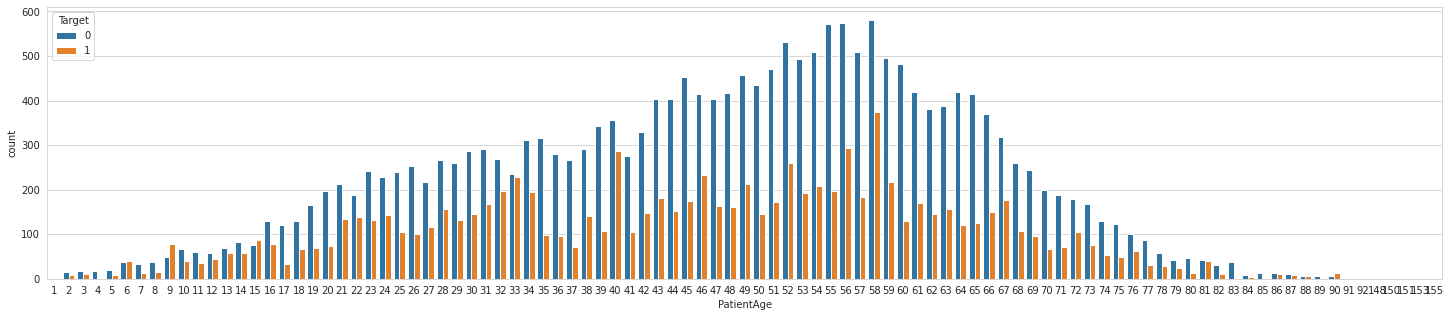
1. **Distribution data over different age group:**

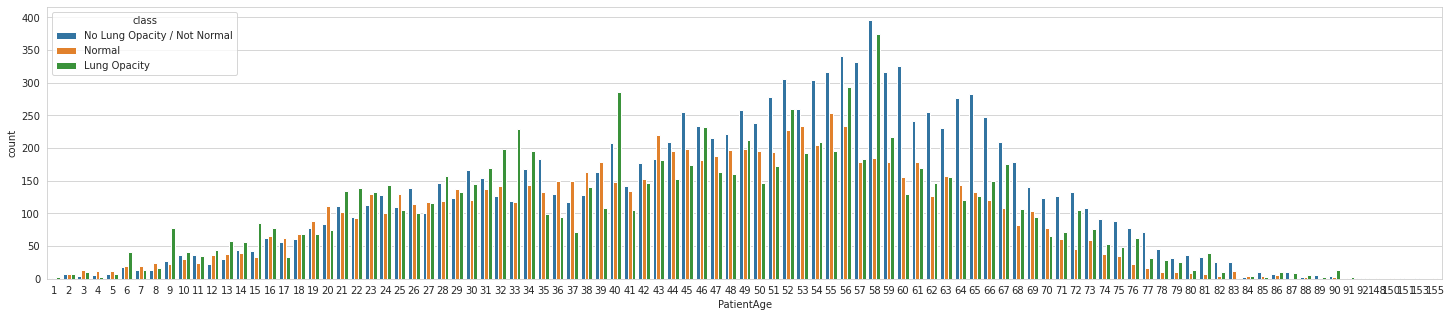
* Three forth of the total test reports falls within age group of 25 – 70 years. The peak is between 50-60 years. It can be observed that lung related ailments are more common in this age group.

This is possibly due to,

* In Adults, the tests conducted are more. Possibly due to adults work out door, consume alcohol, smoke, exercise less (or lazy or inactive than children), suffer from obesity, diabetes, etc. therefore the tests done are more. Not all test result attribute to pneumonia therefore the ratio between pneumonia and normal is more: therefore, the pillars in the below graph.
* In children, the tests for lung ailments conducted are considerably less because of unawareness of possibility of lung related ailments in young age. Therefore, the ratio between Pneumonia to Normal is very less.

* Pneumonia cases are higher between the age group of 30-65 years.
* Possibly due to adults work out door, consume alcohol, smoke, exercise less (or lazy or inactive than children), suffer from obesity, diabetes, etc.
* The lung opacities are spread across the patient age and have peaks within age 50 to 65 years.

****



**Action Taken**: Data augmentation with random sampling is used to balance the data

**Data Preprocessing:**

* **Image scale:** The images have been windowed and leveled already, as have been rescaled to 8-bit encoding and the resolution has been rescaled to (1024, 1024).
* **Data discrepancy:** There is no discrepancy in the data, as the data in the class csv and label csv is the same.
* **Data imbalance:** Data augmentation and random shuffling is used to balance the data.

**Data Augmentation:**

* Given data is not balanced, there are only 21% of total samples are pneumonia and 79% is not pneumonia. Due to this, the model wouldn’t be robust when trained for different data samples (cross validation) and test accuracy was not consistent. Therefore, data augmentation is used while training the model.
* For data augmentations, techniques such as image rotation, scaling, translating, changing brightness, contrast, blurring and sharpening are used.

**Algorithms Used**

Keeping the goal in mind, the below steps are performed to choose a model

* Identify the model that are pretrained. This can reduce the training time for the data
* Check the accuracy and other attributes of the model
* If required Add or drop layers from the model to improve the performance of the model.

The below algorithms are taken into consideration for building the model,

1. VGG - Very Deep Convolutional Networks for Large-Scale Image Recognition. VGG has achieved 92.7% top-5 test accuracy in ImageNet, for a dataset of over 14 million images belonging to 1000 classes
2. Inception – A network known for its speed and accuracy. Due to the large number of training images (~26K), the network is used as backbone network to find the bounding boxes.
3. Mask RCNN - A deep neural network aimed to solve instance segmentation problem based on faster R-CNN. The model generates bounding boxes and segmentation masks for each instance of an object in the image

All the three models are trained and tested with the available datasets by preprocessing the data.

Mask R-CNN gives the highest accuracy among other algorithm and additionally the model gives bounding boxes and class scores for each bounding box, therefore it’s best fit for our problem. Resnet50 is used as backbone network for this model to solve the problem.

1. **Model Building (Step-by-Step) Process**

**Mask R-CNN model is used as final solution to the problem given. Using this algorithm, we are able to predict Pneumonia successfully and also predict the possible areas of lung inflammations on chest radiographs.**

Mask RCNNseparates different objects in an image. Feed it an image, it gives bounding boxes objects, class scores and masks. The model is built as below

1. All the required libraries are imported and the required data sets are included
2. Create configuration parameter object with RESNET50 backbone network, learning rate, batch size, steps per epoch, weight decay, validation steps, number of GPUs, Number of images per GPU, etc.
3. Configure the model by using RESNET50 as the backbone model
4. Define data generator class with loading methods to load pixel data from DICOM image, Mask and Class information. And configure number of classes to be detected by the model.
5. Create model object by passing COCO pre trained weights.
6. Split the training data set into training and validation set
7. Prepare data generators objects for training and validation for model training.
8. Using image augmentation techniques create image augmentation object (used during data generators) to increase the number of training images as identified in EDA. New images are introduced in the training dataset by
   1. Changing the data geometrically by scaling and transforming the images
   2. Changing the brightness and contrast of the images
   3. Blurring or sharpening the Images
9. Start training the model with configuration parameters mentioned in step [B] above.
10. Repeat the steps [B] to [I] for different configuration parameters (hyper parameters and different data size)
11. Fetch weights from model log directory for the best performing model.
12. Create a new model and load the best weights for the model and use it to predict Pneumonia in test data.
13. Calculate evaluation metrics plot confusion matrix and ROC curve
14. Create submission CSV for test data set
15. **Model evaluation**

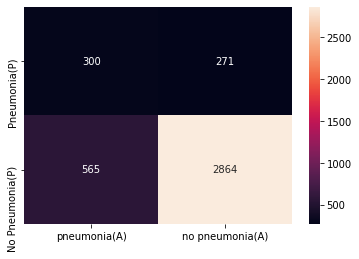
**We are able to predict Pneumonia successfully and also predict the possible areas of lung inflammations on chest radiographs.**

The model identifies bounding boxes for potential lung inflammation and also gives class scores for each of the boxes. When compared with actual label (and bounding box) information, all the predicted bounding boxes weren’t indicating the actual lung inflammations (opacities) therefore a class score threshold of 0.98 is used to categorize the bounding boxes to lung inflammation areas.

Smaller values such as 0.90, 0.95 were adding more number of false positives. Upon evaluating different values 0.98 is finalized as class score threshold for categorization.

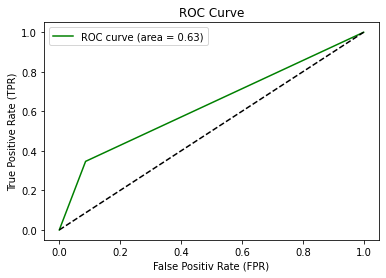
**Model Performance**:

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Accuracy | 79.1% |
| Recall | 0.35 |
| Precision | 0.53 |
| False Negative Rate(Miss Rate) | 0.65 |
| F1 Score | 0.42 |
| True Negative Rate | 0.91 |
| False Positive Rate: | 0.09 |



The model has precision of 0.53 and Recall stands at 0.35. False negatives rate is quite high; therefore, this model can be used at screening stage to predict whether a patient is suffering from Pneumonia but not recommended to use to state a patient not to be having Pneumonia due to high false negatives.

Overall accuracy of the model is 79.1%, however the AUC is 0.63. Therefore, there is 63% chance that model will be able to distinguish between positive class and negative class.



1. **Comparison to benchmark**

**Bench Mark:**

Stanford Machine Learning Group Paper - CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning [[1](https://stanfordmlgroup.github.io/projects/chexnet/)].

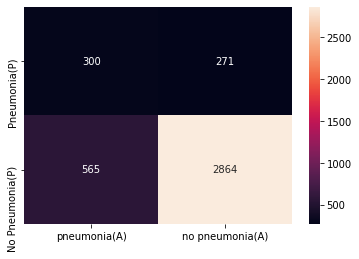
In this paper, F1 score (on an average) came in at 0.435 with 85% with Pneumonia positive (recall).

In comparison, f1-score and recall for our model stand at 0.42 and 51% respectively.

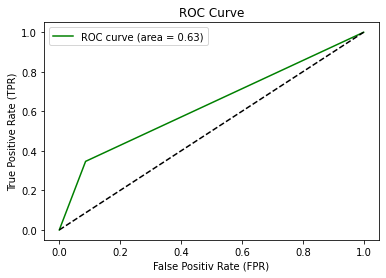
|  |  |  |
| --- | --- | --- |
| **Model** | **F1-Score** | **Recall** |
| CheXNet | 0.435 | 85% |
| Our Model | 0.42 | 51% |

The recall value is lesser as compared to CheXNet although F1-Score is comparable.

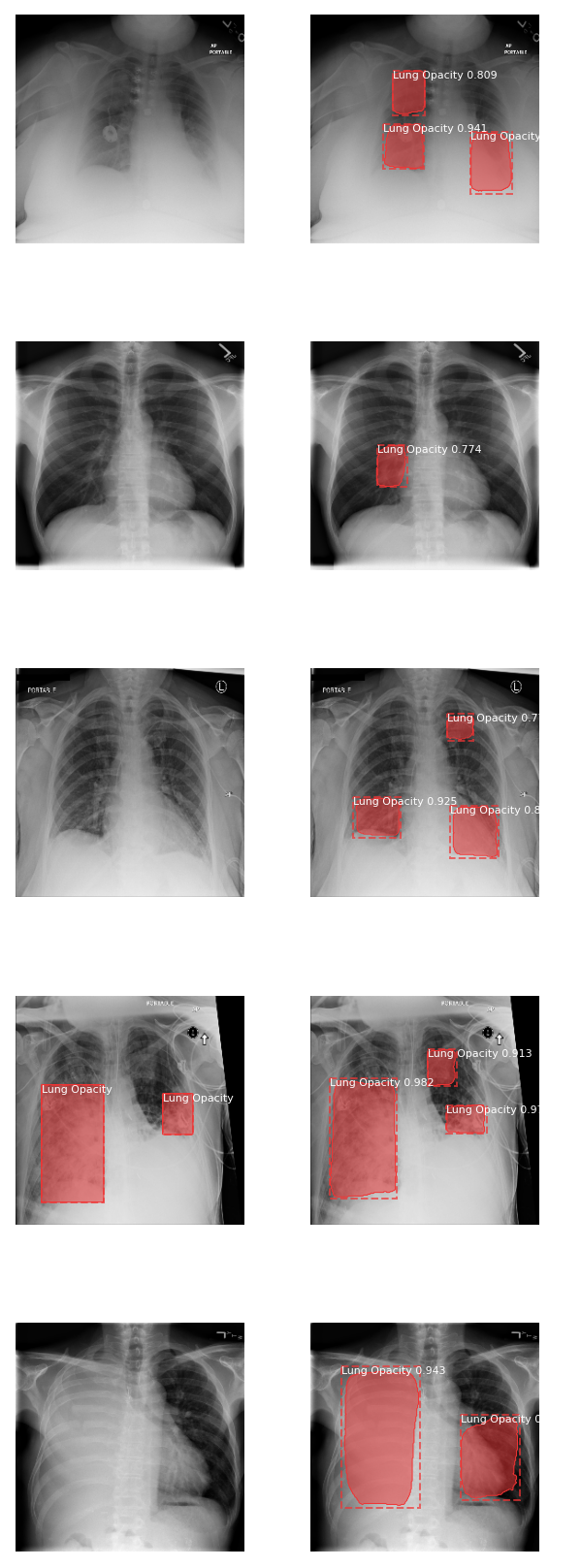
1. **Visualizations**
2. **Confusion Matrix**

****

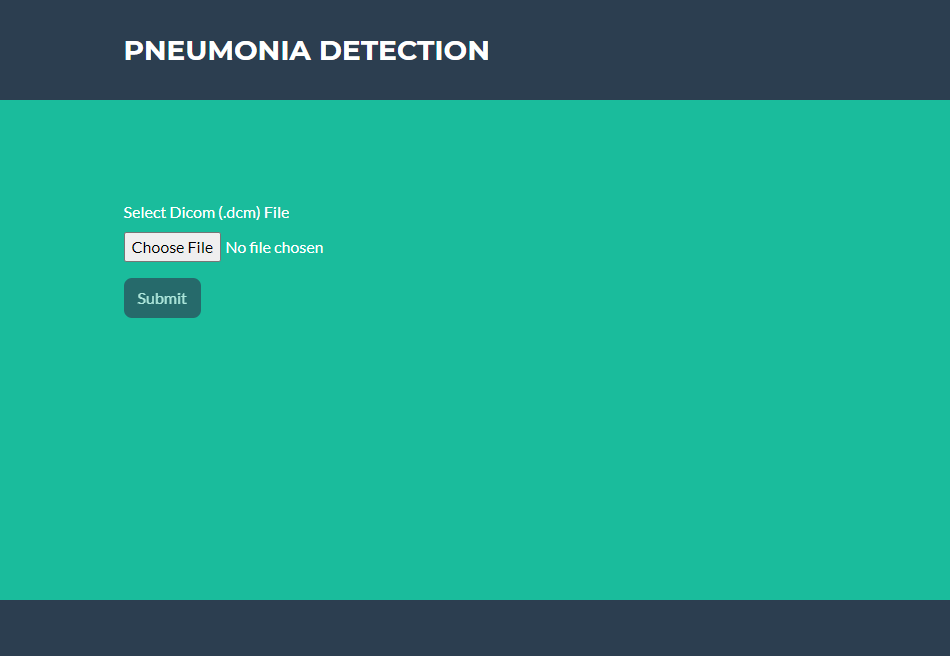
1. **ROC curve**

****

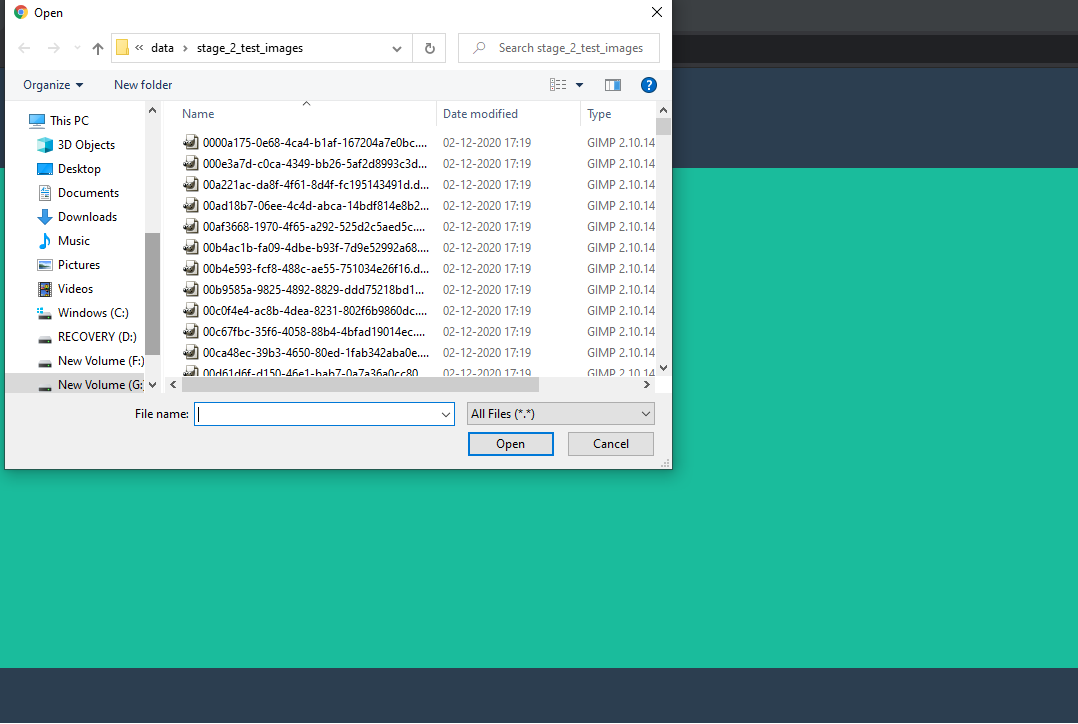
1. **Lung inflammation predicted for sample radiographs**



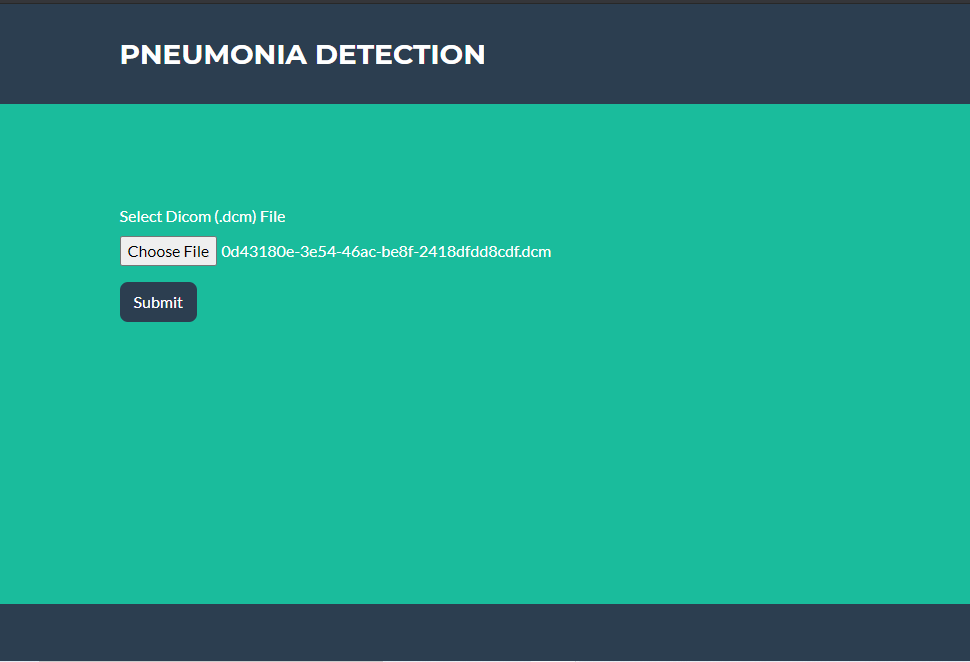
1. **Accessing application on cloud.**
2. Interface to upload chest radiograph



1. Selection of chest radiograph



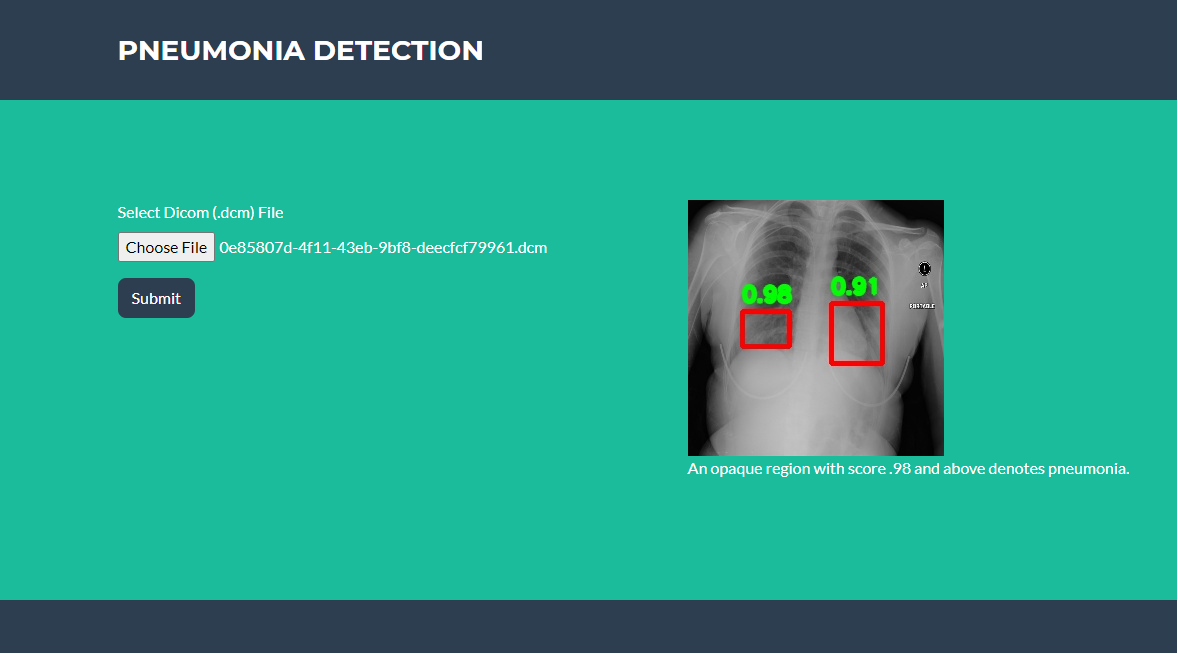
1. Interface to check uploaded file

****

1. Prediction result of Patient 1

****

1. Prediction result of Patient 1

****

1. **Implications**

Due to the moderate accuracy rate and high miss rate, the model is recommended to be used as an assistant at the screening stage to predict whether a patient is suffering from Pneumonia but not recommended to use to state the patient not to be having Pneumonia.

However, the solution has below advantages. Our solution will help in quick decision making (in urgent cases), saves cost, etc.

**High accessibility**: The model is deployed on cloud and a web application is available, therefore it’s easier for the end users to access the application on any computer with browser installed.

**Higher availability:** Diseases, like pneumonia, require early diagnostics for doctors to detect it. False-negative and false-positive results can both be rather destructive. Computer vision excludes the possibility of human error to some degree and serves as an assistant for radiologists. And it’s available 24/7.

**Speed, urgency:** Solution allows doctors to focus more on patients rather than on examining X-ray images. Speed might be crucial for urgent situations.

**Cost savings:** It results in to cost rationalization in terms of lesser staff for examining X-ray images.

1. **Limitations**

* During EDA it is observed that Gender, Age and View Position attributes are having positive correlation with the targets. But, in this model these features are not utilized explicitly.
* Due to high number of false negatives. We recommend this model to be used at screening stage to predict whether a patient is suffering from Pneumonia but not recommended to use to state the patient not to be having Pneumonia.
* Overall accuracy is less.
* At the moment the model is deployed with web application interface, at a time only one chest X-Ray can be uploaded to check the result.
* As an additional feature making a mobile app would have been more useful to the doctors.
* This model could be used as an assistant at the screening stage not as an accurate to replace humans.

1. **Closing Reflections**

During the lifecycle for the project there were many learnings. Such as,

* Model deployment on cloud with web interface
* Exposure to the cloud.
* Dealing with large data sets and running multiple GPUs.
* Using python Flask.
* Using multiple algorithms for the same problem and evaluating the models

Improvements that could be done are,

* Improving the model to make decision accurately on false negatives and increase the overall precision.
* A mobile app for the doctors where they can see the reports for the concerned patients.
* Features to upload as bulk radiographs of patients and see the reports
* Integration of cloud deployed model to their own applications running on their computers.

The software can be commercialized in following ways:

* Hosting as SaaS where customers make accounts and pay for using.
* Licensing software to clients who can deploy it on their infrastructure.

1. **Appendix**

**Code:**

|  |  |  |
| --- | --- | --- |
| **1** | Mask R-CNN Model | <https://github.com/umesh-kushwaha/CapstoneProject-Pneumonia-detection.git> |

1. **References:**
2. [CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning.](https://stanfordmlgroup.github.io/projects/chexnet/)